a. Introduction

The insurance industry is one of the first industries that benefited from the advancement of data science. The companies within this industry that started relying on insights derived from data to make optimal business decisions outperform non-data-driven companies. With the advent of data science, and application of advanced machine learning algorithms, the insurance industry is being transformed by eliminating empirical and intuition-based decisions and replacing them with logical, factual and data-based processes, free of human errors.

In this capstone, it will be demonstrated how one business process can be optimized by application of data analytics. Inferences backed by statistical analysis will be explained in a simplified context by use of data mining and data visualization.

As a short context of one of the business problems that will be given answer in this paper is transaction delays due to risk assessment procedure. House loan financing companies as a subset sector within the insurance industry, rigorously review each loan application to ensure that the applicant has the proven capability to pay off the loan within the desired timetable. If this evaluation process is bypassed and not given the needed weight of importance, the company runs the risk of being ruined by defaulted loans. This evaluation process results in processing and transaction downtime (in days) when the company agents have to individually receive and evaluate eligibility of customers who are applying for house loans. This downtime directly equates to opportunity cost not only to the company, but also to the customer.

By leveraging data mining techniques, this inefficiency may be minimized if not eliminated. A Supervised learning model may be created to automate prediction of loan eligibility of customers based on their provided personal information (eg. marital status, gender, income, etc), lessening if not eliminating the need for manual assessment per loan applicant. The relationships, patterns or trends between customer features (given in the data set) as predictors of loan eligibility may be uncovered. The importance of such relationships, if they are proven to be significant predictors of eligibility, cannot be overemphasized as this will streamline the loan eligibility procedure to maximize productivity.

Along with this, other business insights may also be determined such as which customer segments are best targeted for the marketing campaigns. This in effect will help redirect marketing costs to proper targets, with the goal of increasing insurance sales, and thus maximizing possible profit returns.

b. Background and Objective

House loan insurance company, ALPHA Financing, provided the loan applicants information datasets for this study. These datasets will be used to attain the specific objective/s as iterated by ALPHA Financing for this project:

Objective 1. Create a prediction model

- This model should be proven to be of significant accuracy. Only then it will be able to help automate eligibility feedback to house loan applicants who submitted their application via the online form on their website.
- A particular attention should be given to the model prediction error, *false positives*, or those customers who were classified as eligible applicants but actually are ineligible, as this type of error will be more costly. *False negative* predictions will cause the company to miss potentially good customers, but false positives will cause financial loss. The former is equivalent to loss of potential monetary income, but the latter is loss of actual money. Both errors are undesirable, and therefore should both be minimized, but false negative is more tolerable than the other.

Objective 1.a Determine the best customer segment for house loan applicants

- To satisfy Objective (1), the best customer fit must first be determined as this will be beneficial to help boost the marketing efforts of the company. Knowing where and who to target may maximize marketing budgets.

Only the customer details provided for this study are considered, and are treated as factual, but may be biased which means that the model developed may be a good fit for an existing dataset but not for new data. Usage of confusion matrix will be applied to lessen the effect of bias on the model. Accuracy of prediction and analysis may be limited by the truthfulness of data supplied, and other factors that are not part of the customer information sheet being filled up for house loan application.

b.1 Definition of Terms:

Customer - loan applicant

Loan eligibility procedure - the process of cross checking customer information to predetermined standards before allowing the customer to continue with the process of loan application

The data utilized in this project study is that of customer information relevant to the standards of loan eligibility evaluation process.

The company provided the training and testing datasets in .xslx format containing the following customer information:

Variable	**Description**
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Undergraduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	(Target) Loan approved (Y/N)